Alignment-Enriched Tuning for Patch-Level Pre-trained Document Image Models

Lei Wang[1,2], **Jiabang He**[1], **Xing Xu**[1*], **Ning Liu**[3], **Hui Liu**[4]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
[2] Singapore Management University, Singapore
[3] Beijing Forestry University, China
[4] Beijing Rongda Technology Co., Ltd., China
demolwang@gmail.com, JiaBangH@outlook.com, xing.xu@uestc.edu.cn
liuning0928@bjfu.edu.cn, ryuki122382@gmail.com

Abstract

Alignment between image and text has shown promising improvements on patch-level pre-trained document image models. However, investigating more effective or finer-grained alignment techniques during pre-training requires a large amount of computation cost and time. Thus, a question naturally arises: *Could we fine-tune the pre-trained models adaptive to downstream tasks with alignment objectives and achieve comparable or better performance?* In this paper, we propose a new model architecture with alignment-enriched tuning (dubbed AETNet) upon pre-trained document image models, to adapt downstream tasks with the joint task-specific supervised and alignment-aware contrastive objective. Specifically, we introduce an extra visual transformer as the alignment-ware image encoder and an extra text transformer as the alignment-ware text encoder before multimodal fusion. We consider alignment in the following three aspects: 1) document-level alignment by leveraging the cross-modal and intra-modal contrastive loss; 2) global-local alignment for modeling localized and structural information in document images; and 3) local-level alignment for more accurate patch-level information. Experiments on various downstream tasks show that AETNet can achieve state-of-the-art performance on various downstream tasks. Notably, AETNet consistently outperforms state-of-the-art pre-trained models, such as LayoutLMv3 with fine-tuning techniques, on three different downstream tasks. Code is available at https://github.com/MAEHCM/AET.

Introduction

Self-pretraining techniques aiming to learn generic representations have recently proved to be highly effective for document image understanding. Notably, transfer learning based on pre-trained document image models yields strong performance on various document related downstream tasks (Xu et al. 2020, 2021b; Garncarek et al. 2021; Hong et al. 2022; Wu et al. 2021; Li et al. 2021a,c,d; Lee et al. 2022; Huang et al. 2022; Li et al. 2022). A typical self-pretraining method in document image understanding is to pre-train a model on a large amount of pairs of the document images and OCR texts with layout information underlying the constraints of unsupervised losses. Downstream tasks then usually leverage the pre-trained weights for initialization. After that, the initialized model is fine-tuned for a downstream task using a task-specific supervised objective.

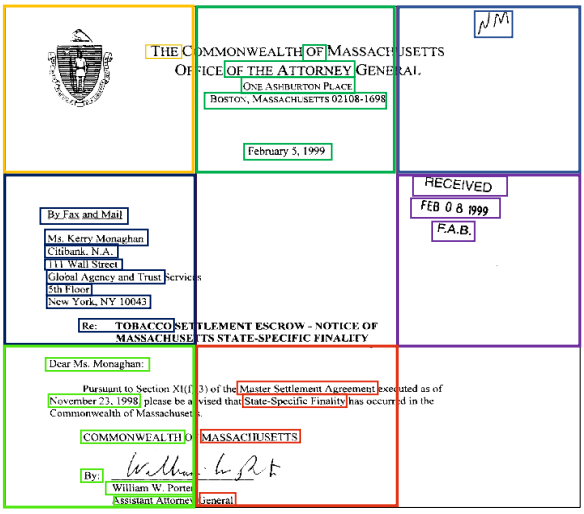

Figure 1: A document image example containing rich contextual text and layout information.

As shown in Figure 1, a document image contains rich contextual text and structural information, requiring fine-grained interaction modeling between image and text. Although self-supervision achieves large progress in document image related-tasks, most existing pre-trained document image models are trained with coarse self-supervised losses, which ignore fine-grained interaction modeling between image and text. Inspired by this, DocFormer (Appalaraju et al. 2021) learns to reconstruct image pixels through a CNN decoder, SelfDoc (Li et al. 2021c) proposes to regress the masked region feature, and the latest model LayoutLMv3 (Huang et al. 2022) introduces a word-patch alignment objective by reconstructing masked patch tokens of the visual modality. Although alignment between image and text has shown promising improvements on large-scale self-supervised pre-trained document image models, investigating more effective or finer-grained alignment techniques during the pre-training requires huge computational cost and time. Thus, a question naturally arises: *Could we fine-tune the pre-trained models*

*Corresponding Author

adaptive to downstream tasks with more alignment objectives and achieve comparable or better performance?

In recent few years, with the wide success of large-scale pre-trained models, a range of tuning techniques has arisen to adapt these general-purpose models to downstream tasks (Howard and Ruder 2018; Jiang et al. 2019; Houlsby et al. 2019; Gururangan et al. 2020; Xu et al. 2021a; Liu et al. 2021a,b; Li and Liang 2021). The model tuning strategy (i.e., fine-tuning) (Howard and Ruder 2018) tunes all model parameters during adaptation. To improve training efficiency, adapters (Houlsby et al. 2019) and prompt tuning (Li and Liang 2021; Liu et al. 2021b,a) tune pre-trained models by a trivial amount of parameters, but they may suffer a small performance degradation compared with fine-tuning. In addition, task-adaptive pre-training (Gururangan et al. 2020) continues to train the pre-trained models with the unlabeled data of specific tasks. However, it is inconvenient to incorporate alignment loss into existing tuning strategies due to the limitation of fixed model architectures of pre-trained models.

To enable the pre-trained models to have the ability of modeling alignment, ALBEF (Li et al. 2021b) and TCL (Yang et al. 2022) encode the document image and text independently with an extra image encoder and an extra text encoder before fusion by a multimodal encoder. The model is then pre-trained with alignment-aware strategies to improve the learned features' expressiveness, which is essential for joint multi-modal representation learning. However, as mentioned before, investigating more effective or finer-grained alignment techniques during the pre-training stage requires much computation cost and time.

To bridge the above research gap, in this paper, we propose a new model architecture with alignment-enriched tuning (termed AETNet), which tunes pre-trained document image models adaptively to enable downstream tasks with the joint task-specific supervised and alignment-aware contrastive objective. As shown in Figure 2, compared with fine-tuning and prompt-tuning, in our proposed AETNet, we introduce an extra visual transformer as the alignment-ware image encoder and an extra text transformer as the alignment-ware text encoder before multimodal fusion. We consider alignment in the following three aspects: 1) document-level alignment by leveraging the cross-modal and intra-modal contrastive loss; 2) global-local alignment for modeling localized and structural information in document images; and 3) local-level alignment for more accurate patch-level information.

We evaluate our AETNet method on various downstream document image understanding tasks, including FUNSD (Jaume, Ekenel, and Thiran 2019) for form understanding, CORD (Park et al. 2019) for receipt Understanding, DocVQA (Mathew, Karatzas, and Jawahar 2021) for document visual question answering, and a sampled subset RVL-CDIP-1 from RVL-CDIP (Harley, Ufkes, and Derpanis 2015) for document image classification. In terms of performance, the proposed AETNet method consistently outperforms existing state-of-the-art pre-trained models with fine-tuning on various downstream tasks. Notably, with the help of alignment-enriched tuning, AETNet achieves better performance compared with general fine-tuning and prompt tuning strategies. We also carry out ablation studies with detailed analysis to investigate the effectiveness of each alignment loss in AETNet. Lastly, we conduct case studies over real examples from the FUNSD task to show that AETNet benefits from the alignment modeling.

Related Work

Multimodal Pre-training Multimodal self-supervised pre-training has been successfully applied in document images through effectively leveraging image, layout, contextual text information (Xu et al. 2020, 2021b; Garncarek et al. 2021; Powalski et al. 2021; Wu et al. 2021; Li et al. 2021a,c; Appalaraju et al. 2021; Li et al. 2021d; Wang, Jin, and Ding 2022; Gu et al. 2022; Kim et al. 2022). LayoutLM and its following works consider the layout information as a type of two-dimensional positional vectors and fuse their transformed vectors with text embeddings for the multimodal pre-trained model (Xu et al. 2020; Li et al. 2021a; Hong et al. 2022; Lee et al. 2022). Some works extract CNN grid features (Xu et al. 2021b; Appalaraju et al. 2021) and some (Xu et al. 2020; Powalski et al. 2021; Li et al. 2021c; Gu et al. 2021) rely on object detectors to extract region features. However, these works are either limited by heavy computation bottleneck or require region supervision. Recently researchers make many efforts to overcome the above limitations caused by CNN. Inspired by Vision Transformer (ViT) (Dosovitskiy et al. 2021), most rely on separate self-attention networks to learn visual features to reduce computational cost (Xue et al. 2021; Li et al. 2021b; Dou et al. 2021). For instance, ViLT (Kim, Son, and Kim 2021), one of works in vision-and-language pre-training utilizing ViT, learns visual features with a lightweight linear layer and significantly reduce the model parameters and running time. Following ViLT, LayoutLMv3 (Huang et al. 2022) is the first work to take advantage of patch-level embeddings without CNNs in Document images. For modeling fine-grained interaction and alignment between document image and their OCR text, we introduce alignment-aware ViT (Dosovitskiy et al. 2021) and RoBERRTa (Liu et al. 2019) encoders before fusion.

Tuning Techniques Fine-tuning is the useful paradigm for tuning large pretrained language models for downstream tasks (Devlin et al. 2019). In recent few years, different fine-tuning techniques have been proposed. Model tuning (Howard and Ruder 2018) requires tuning the entire pre-trained model for each downstream task. Mixout (Lee, Cho, and Kang 2019) randomly replaces part of the model parameters with the pre-trained weights during fine-tuning. Child-Tuning (Xu et al. 2021a) updates parameters within the child network via a gradient mask. There are also studies focusing on parameter-efficient fine-tuning, such as adapter-based methods (Houlsby et al. 2019), and the prompt tuning methods (Li and Liang 2021; Liu et al. 2021b,a). These works tune pre-trained models by a trivial amount of parameters while they may suffer a small performance degradation compared with fine-tuning. Beyond, it is inconvenient to incorporate alignment loss into these existing tuning strategies due to the limitation of fixed model architectures of pre-trained models.

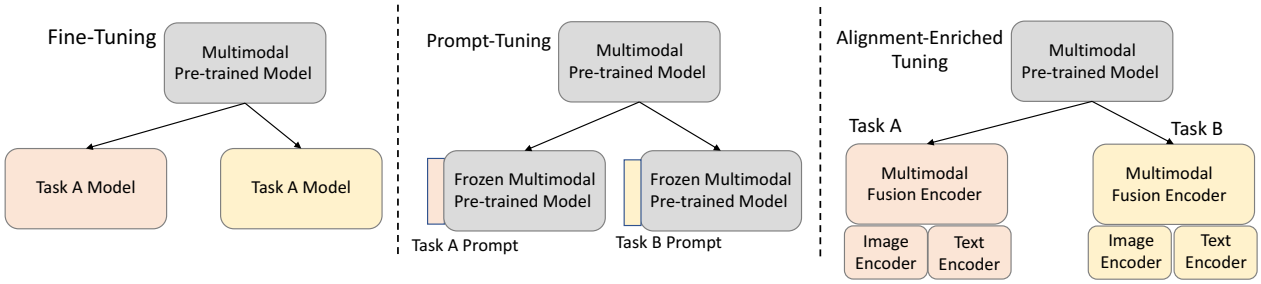

Figure 2: Fine-tuning requires tuning the entire pre-trained model for each downstream task. Prompt-tuning only requires tuning a trivial amount of extra parameters for each downstream task. Alignment-enriched tuning adds an extra alignment-aware image encoder and an extra alignment-aware text encoder before the multimodal pre-trained model. Alignment-enriched tuning trains the entire parameters together for each downstream task.

Methodology

In this section, we present a new model architecture with the proposed alignment-enriched tuning (AETNet) for the transfer of the patch-level pre-trained models in document images to downstream tasks. We first introduce the model architecture, followed by the tuning process. Then we detail the tuning objectives including downstream task specific supervised objective, document-level image-text contrastive objective, intra-modal contrastive objective, global-local image-text contrastive objective, and patch-level image-text alignment objective.

Model Architecture

The model architecture of our method AETNet is shown in Figure 3. The model architecture consists of a large-scale patch-level multimodal pre-trained model, an extra alignment-aware image encoder, and an extra alignment-aware text encoder. For the multimodal pre-trained model, we employ LayoutLMv3 (Huang et al. 2022) as our pre-trained model because of not only its state-of-the-art performance on downstream tasks but also its patch-level image encoder that is suitable for the AETNet process. We use a RoBERTa (Liu et al. 2019) as the extra alignment-aware text encoder. Likewise, we use a ViT (Dosovitskiy et al. 2021) as the extra alignment-aware image encoder, which is initialized using weights pre-trained on ImageNet-1k from DeiT (Touvron et al. 2021). The implementation details of extra text and image encoders, are described in Experiment Section.

Tuning Process

In the downstream tuning process, our model is trained with the joint loss of downstream task specific objective and the proposed alignment-enriched objective. As shown in Figure 3, an input document image I is first encoded into a sequence of patch-level visual representation vectors: $\{v_{cls}, v_1, ..., v_N\}$ through the alignment-aware ViT, where cls denotes the special token [CLS] and N is the number of patches. Meanwhile, the input text (i.e., the textual content of this document image, which is obtained by the open-source OCR toolkit, Tesseract is fed into the alignment-aware RoBERTa to be transformed into a sequence of text token representation vectors: $\{w_{cls}, w_1, ..., w_L\}$, where L is the length of the input text tokens. Before fusion with LayoutLMv3, we compute alignment-aware losses based on the obtained representations in terms of alignment-aware objectives to let the obtained representations preserve alignment information. Then, we derive the patch-level input image embeddings for LayoutLMv3 by the sum of obtained alignment-enriched image representation vectors and original input image embeddings of LayoutLMv3. Likewise, input text embeddings for LayoutLMv3 are derived by the sum of obtained alignment-enriched text representation vectors and original input text embeddings of LayoutLMv3. Lastly, LayoutLMv3 transforms the newly fused embeddings into output hidden representation for computing supervised task-specific loss.

Tuning Objectives

The full tuning objective of AETNet consists of five objectives: the downstream task-specific subjective objective (SO), global-level cross-modal alignment (GCMA), global-level intra-modal contrastive (GIMC), Global-Local MI Maximization (GLMI), and Local-level Cross-Modal Alignment (LLCMA), shown as below:

$$\mathcal{L}_{aet} = \mathcal{L}_{so} + \mathcal{L}_{ditc} + \mathcal{L}_{imc} + \mathcal{L}_{glitc} + \mathcal{L}_{pita}. \quad (1)$$

In the following, we elaborate each objective in details.

Supervised Objective (SO) In the vanilla fine-tuning framework, the model is first initialized with the parameters of large pre-trained models (LPM). Then, the initialized model is fine-tuned for a certain downstream task using the task-specific objective on the corresponding dataset. Unlike vanilla fine-tuning, our model consists of three modules: LPM, the alignment-aware image encoder, and the alignment-aware text encoder. We train these three modules together in our model with the task-specific objective: $\mathcal{L}_{so}(I, T, Y)$, where image I and OCR text T are the input and Y is for the ground-truth labels. We take the semantic entity labeling task (Jaume, Ekenel, and Thiran 2019; Park et al. 2019), one of the downstream tasks mentioned in Experiment Section, as an example. L_{so} is the cross entropy loss based on OCR tokens' predictions and their corresponding ground-truth labels.

Document-Level Image-Text Contrastive (DITC) The purpose of DITC is to learn better document(global)-level alignment cross-modality through a contrastive

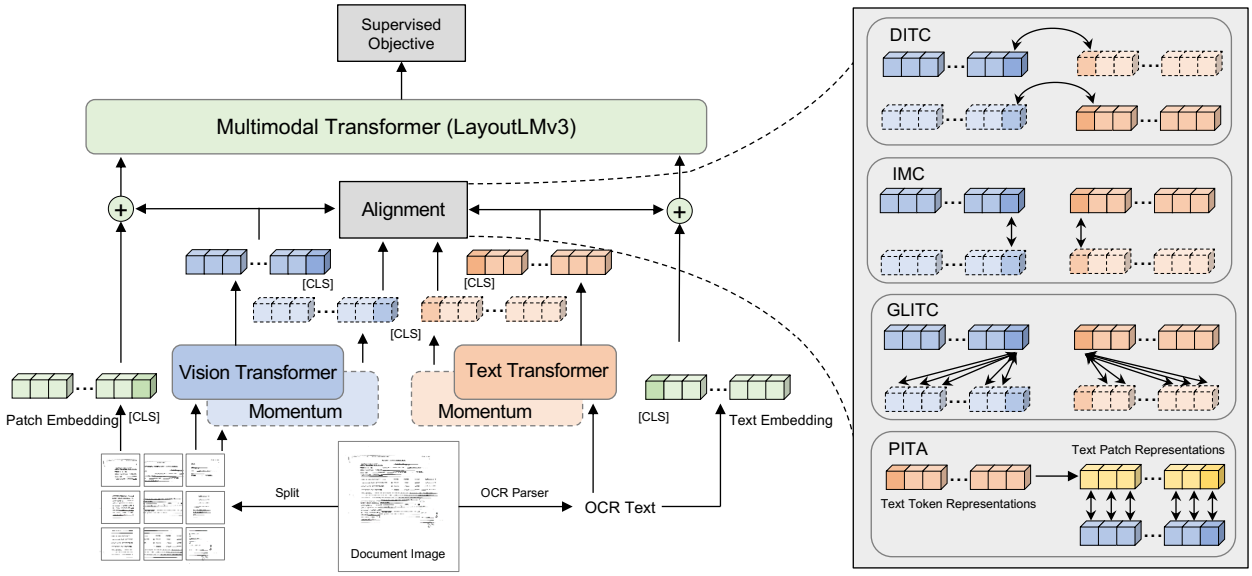

Figure 3: An overview of the framework of AETNet. The model architecture consists of a vision transformer, a text transformer, and a fusion encoder. Image and text encoders have their paired momentum encoder updated by the momentum-based moving average. The alignment objective contains four contrastive losses (i.e., DITC, IMC, GLITC and PITA) for both alignment-enriched cross-modal and intra-modal representation learning.

loss, which pulls representations of matched image-text pairs to be close and pushes the unmatched ones apart. To model global information, we apply similarity functions $sim(v_{cls}, \tilde{w}_{cls}) = f_{img}(v_{cls})^\top f_{txt}(\tilde{w}_{cls})$ and $sim(w_{cls}, \tilde{v}_{cls}) = f_{text}(w_{cls})^\top f_{img}(\tilde{v}_{cls})$ for the DITC loss, where f_{img} and f_{txt} are two projection heads for mapping hidden vectors into low-dimensional representations in the contrasitve loss space and v_{cls} and w_{cls} are two [CLS] vectors from alignment-aware image and text encoders, respectively. Next, we refer to MoCo (He et al. 2020) and ALBEF (Li et al. 2021b) to maintain two memory queues (i.e., $\tilde{I}_{cls} = \{\tilde{v}^1_{cls}, ..., \tilde{v}^K_{cls}\}$ for images, $\tilde{T}_{cls} = \{\tilde{w}^1_{cls}, ..., \tilde{w}^K_{cls}\}$ for text) to record the most recent K image and text representations from the momentum alignment-aware encoders. The document-level contrastive loss for a pair of document image I and OCR text T can be defined as follows:

$$\mathcal{L}_{cl}(v_{cls}, \tilde{w}_{cls}, \tilde{T}_{cls}) = -\log \frac{\exp(sim(v_{cls}, \tilde{w}_{cls})/\tau)}{\sum_{k=1}^{K} \exp(sim(v_{cls}, \tilde{w}^k_{cls})/\tau)}, \tag{2}$$

where τ is the learnable temperature rate. Considering image-to-text and text-to-image together, DITC loss is defined as:

$$\mathcal{L}_{ditc} = \frac{1}{2}\mathbb{E}_{(I,T)\sim B}\big[\mathcal{L}_{cl}(v_{cls}, \tilde{w}_{cls}, \tilde{T}_{cls}) + \mathcal{L}_{cl}(w_{cls}, \tilde{v}_{cls}, \tilde{I}_{cls})\big], \tag{3}$$

where B is a batch of image-text pairs. In addition, We follow ALBEF (Li et al. 2021b) to guide the training of DITC loss by pseudo-targets generated by the momentum model.

Intra-Modal Contrastive (IMC) The goal of IMC is to learn more accurate representations within the same modality. For the visual modality, we generate positive latent image representation $\{v^+_{cls}, v^+_1, ..., v^+_N\}$ for the anchor augmented image I^+ by feeding the anchor image into the momentum alignment-aware image encoder. For the text modality, the positive latent representation $\{w^+_{cls}, w^+_1, ..., w^+_L\}$ for the anchor text T is produced by the momentum model in the same way as image samples. Two momentum queues used for the DITC loss also provide negative samples for the IMC loss. The IMC loss is define as follows:

$$\mathcal{L}_{imc} = \frac{1}{2}\mathbb{E}_{(I,T)\sim B}\big[\mathcal{L}_{cl}(v_{cls}, v^+_{cls}, \tilde{I}_{cls}) + \mathcal{L}_{cl}(w_{cls}, w^+_{cls}, \tilde{T}_{cls})\big]. \tag{4}$$

Global-Local Image-Text Contrastive (GLITC) Inspired by TCL (Yang et al. 2022), we introduce the GLITC loss to capture localized and structural information within the single modality by modeling interactions between the document-level representation and local regions. Specifically, for the visual modality, we use global representation v_{cls} and compute the contrastive loss with momentum image patch representations$\{\widetilde{v}_2, \ldots, \widetilde{v}_N\}$. Likewise, for the the text modality, we use the text [CLS] representation w_{cls} and momentum text token representations $\{\widetilde{w}_2, \ldots, \widetilde{w}_L\}$ for computing loss. We encourage to model interactions between global information and lcoal information within the same modality by minimizing the following two-side interaction contrastive loss:

$$\mathcal{L}_{glitc} = \frac{1}{2}\mathbb{E}_{(I,T)\sim B}\big[\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{cl}(v_{cls}, \tilde{v}_i, I^-_B) + \frac{1}{L}\sum_{j=1}^{L}\mathcal{L}_{cl}(w_{cls}, \tilde{w}_j, T^-_B)\big], \tag{5}$$

where I_B^- and T_B^- are negative image patch and text token representation vectors, respectively. These negatives are from other data examples in the same batch B.

Patch-Level Image-Text Alignment (PITA) Although the introduced DPITC loss has the ability to capture certain cross-modal localized and structural information in the input, it ignores patch-level alignment between patch-level images and patch-level contextual text. Patch-level alignment is critical to learning more accurate and finer-grained representations for more accurate document understanding. Therefore, we propose a novel patch-level alignment loss as a complement to the previous three contrastive losses. As shown in Figure 1, to compute the patch-level alignment loss, we first find text tokens matching image patches by a rule-based strategy that locates text in the image patches by layout information obtained from the OCR toolkit. Then, we average the representation vectors of the matched text tokens to derive patch-level text representations for the matched text $\{t_1, t_2, ..., t_N\}$. The PITA loss is defined below:

$$\mathcal{L}_{pita} = -\frac{1}{N}\Big(\sum_{i=1}^{N} \frac{\langle v_i, t_i \rangle}{||v_i|| \cdot ||t_i||}\Big). \tag{6}$$

Experiment

Experiment Setting

Model Configures Since we aim to investigate the impact of the proposed fine-grained alignment losses during the tuning process, we thus adopt the patch-level multimodal pre-trained model (LayoutLMv3 (Huang et al. 2022)) instead of document-level pretrained models as our fusion model. LayoutLMv3 is pre-trained on 11 million document images of a large IIT-CDIP (Harley, Ufkes, and Derpanis 2015) dataset. LayoutLMv3$_{\text{BASE}}$ uses a 12-layer Transformer (Vaswani et al. 2017) Encoder with 12 self attention heads, 768 hidden size, and 3,072 intermediate size for feed-forward network. LayoutLMv3$_{\text{BASE}}$ uses a 24-layer Transformer (Vaswani et al. 2017) Encoder with 16 self attention heads, 1,024 hidden size, and 4,096 intermediate size for feed-forward network. Patch size in LayoutLMv3 is 16. The alignment-aware ViT and RoBERTa encoders before fusion are initialized with the model parameters released by DeiT (Touvron et al. 2021) and RoBERTa (Liu et al. 2019), respectively. The size of the momentum queues of all encoders is 65536.

Downstream Tasks We evaluate the proposed alignment-enriched tuning (AET) method on three document multimodal downstream tasks as follows:

Form and Receipt Understanding is a sequential labeling task, which aims to assign a label to each word. Form and receipt understanding requires an understanding of visual information and textual content extracted from structural forms in document images. We follow fine-tuning settings in LayoutLMv3 to evaluate AET on two public datasets, i.e., FUNSD (Jaume, Ekenel, and Thiran 2019) and CORD (Park et al. 2019). FUNSD is a dataset sampled from the RVL-CDIP dataset (Harley, Ufkes, and Derpanis 2015) about noisy scanned form understanding. It consists of 199 documents (149 documents for training and 50 documents for testing) and 9,707 semantic entities. CORD is a receipt key information extraction dataset, including 1,000 receipts and 30 semantic labels defined under 4 categories, where 800 samples are used for training, 100 for validation, and 100 for testing.

Model	FUNSD F1↑	CORD F1↑
BERT$_{base}$ (Devlin et al. 2019)	60.26	89.68
RoBERTa$_{base}$ (Liu et al. 2019)	66.48	93.54
BROS$_{base}$ (Hong et al. 2022)	83.05	95.73
UDoc (Gu et al. 2021)	87.93	98.94†
LayoutLMv2$_{base}$ (Xu et al. 2021b)	82.76	94.95
DocFormer$_{base}$ (Appalaraju et al. 2021)	83.34	96.33
LayoutLMv3$_{base}$ (Huang et al. 2022)	89.82	95.97
AETNet$_{base}$ (Ours)	**91.55**	**97.04**
BERT$_{large}$ (Devlin et al. 2019)	65.63	90.25
RoBERTa$_{large}$ (Liu et al. 2019)	70.72	93.80
BROS$_{large}$ (Hong et al. 2022)	84.52	97.40
LayoutLMv2$_{large}$ (Xu et al. 2021b)	84.20	96.01
DocFormer$_{large}$ (Appalaraju et al. 2021)	84.55	96.99
LayoutLMv3$_{large}$ (Huang et al. 2022)	90.94	97.01
AETNet$_{large}$ (Ours)	**92.33**	**97.52**

Table 1: Performance comparison with existing published pre-trained models with fine-tuning on FUNSD and CORD datasets. The results of LayoutLMv3 are based on our implementations with the released UniLM (Dong et al. 2019). The score† is not directly comparable to other scores.

Document Visual Question Answering is to predict the answer given an document image and a question, which requires models understanding the knowledge in documents and learning to reason over documents to answer. We follow the task formulation in LayoutLMv3, considering this task as an extractive QA problem. Specifically, the model is required predicting start and end positions of the answer in the document. This is a binary classification over each text token. We follow the official partition of the DocVQA (Mathew, Karatzas, and Jawahar 2021) dataset, which consists of 10,194/1,286/1,287 images with 39,463/5,349/5,188 questions for training/validation/test, respectively.

Document Image Classification is a document classification task aiming to predict the category of a given document, which requires of an understanding of document contents. Due to the limitations of our servers, we sampled a subset from RVL-CDIP (Harley, Ufkes, and Derpanis 2015), termed RVL-CDIP-1 , including 10,000 data examples with 16 categories from the RVL-CDIP datasets of 400,000 document images as our evaluation dataset. RVL-CDIP-1 is divided into 8000 training samples, 1000 validation samples, and 1000 test samples.

AETNet on Downstream Tasks

The detailed description of hyper-parameters, including running epochs, learning rate, batch size, and optimizer, for our method on three downstream tasks and four datasets, are referred to https://github.com/MAEHCM/AET.

Evaluation on Form and Receipt Understanding We compare the proposed model architecture with alignment-

Model	DocVQA ANLS↑
BERT$_{base}$ (Devlin et al. 2019)	63.72
RoBERTa$_{base}$ (Liu et al. 2019)	66.42
LayoutLMv2$_{base}$ (Xu et al. 2021b)	78.08
LayoutLMv3$_{base}$ (Huang et al. 2022)	78.76
AETNet$_{base}$ (Ours)	**79.73**

Table 2: Performance comparison on DocVQA dataset.

enriched tuning (AETNet) to three types of self-supervised pre-training approaches using fine-tuning: (1) Pre-trained models only using text modality. BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) is based on Transformer architecture and only use text information; (2) Pre-trained models with text and layout modalities. BROS (Hong et al. 2022) incorporates layout information by encoding relative layout positions; (3) Pre-trained models with text, layout and image modalities. UDoc (Gu et al. 2021) adopts object proposals from document images and concatenate region features and text embeddings. LayoutLMv2 (Xu et al. 2021b) use a CNN network to extract image features and then feed it with layout and text information into multimodal Transformer. DocFormer (Appalaraju et al. 2021) extract image features with CNN. LayoutLMv3 (Huang et al. 2022) replaces CNN backbones with patch-level linear embedding layers.

Table 1 reports the results of the comparison AETNet to fine-tuning for existing published pre-trained models on FUNSD and CORD datasets. Our APTNet achieves state-of-the-art performance while outperforming the previous fine-tuning based methods. For transferring the pre-trained model to the downstream tasks, we compared AETNet$_{base}$ to the reported fine-tuning for LayoutLMv3$_{base}$ (Huang et al. 2022) and our implemented fine-tuning for LayoutLMv3$_{base}$. AETNet$_{base}$ improves $+1.25$ and $+1.73$ on FUNSD and $+0.51$ and $+1.07$ on CORD, revealing the necessity of conducting document-level and patch-level intra-modal and cross-modal alignment after pre-training.

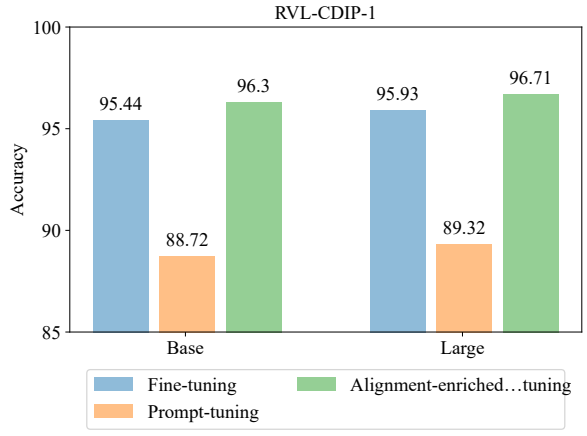

Figure 4: Performance comparison on the sampled subset RVL-CDIP-1 dataset.

#Pre-train Models	Training tasks	FUNSD F1↑	CORD F1↑
LayoutLMv3$_{base}$	Fine Tuning	89.82	95.97
	P-Tuning (Liu et al. 2021a)	83.75	87.37
	AETNet	**91.55**	**97.04**

Table 3: Comparison to different adaptation approaches on FUNSD and CORD datasets. Fine Tuning: All pre-trained parameters are tuned with the supervised objective. P-Tuning: Prompt parameters of each layer are tuned with the supervised objective. AETNet: Pre-trained parameters and extra alignment-aware image and text encoders are tuned with the supervised objective and our proposed alignment loss. All results are based on our implementation.

Evaluation on Document Visual Question Answering Table 2 reports the performance comparison on the DocVQA dataset. We compare AETNet$_{base}$ to methods only requiring text modality, i.e., BERT$_{base}$, RoBERTa$_{base}$, and methods requiring image and text as inputs, i.e., LayoutLMv2$_{base}$ (Xu et al. 2021b) and LayoutLMv3$_{base}$ (Huang et al. 2022). AETNet$_{base}$ substantially outperforms existing baseline methods, achieving an absolute accuracy of $+0.97$ boost compared to LayoutLMv3$_{base}$.

Evaluation on Document Image Classification Figure 4 shows the performance comparison on the sampled subset classification dataset, RVL-CDIP-1. We compare AETNet with the general fine-tuning and prompt-tuning strategies for LayoutLMv3 in both base and large model sizes. Although prompt-tuning is more parameter efficient, it cannot outperform fine-tuning and alignment-enriched tuning. Likewise, fine-tuning outperforms prompt tuning but is worse than alignment-enriched tuning. Overall, in terms of performance, AETNet can achieve desirable results on document image classification with the help of alignment modeling.

Intrinsic Analysis

AETNet *v.s.* Other Tuning Methods As shown in Table 3, we compare AETNet with fine-tuning based LayoutLMv3$_{base}$ and P-tuning based LayoutLMv3$_{base}$ on FUNSD and CORD datasets. With the help of alignment-

Tuning Objective	FUNSD F1↑	CORD F1↑
LayoutLMv3$_{base}$ (Fine-tuning)	89.82	95.97
AETNet$_{base}$ (w/ Supervised Objective)	89.77	96.30
+ DITC	90.44	96.48
+ IMC	90.45	96.61
+ GLITC	90.18	96.37
+ PITA	90.69	96.78
AETNet$_{base}$	**91.55**	**97.04**

Table 4: Ablation study of each loss in the AETNet on FUNSD and CORD datasets. The F1 is reported. DITC, IMC, DLITC, and PITA means Document-Level Image-Text Contrastive, Intra-Modal Contrastive, Global-Local Image-Text, and Patch-Level Image-Text Alignment, respectively. All results are obtained by our implementation.

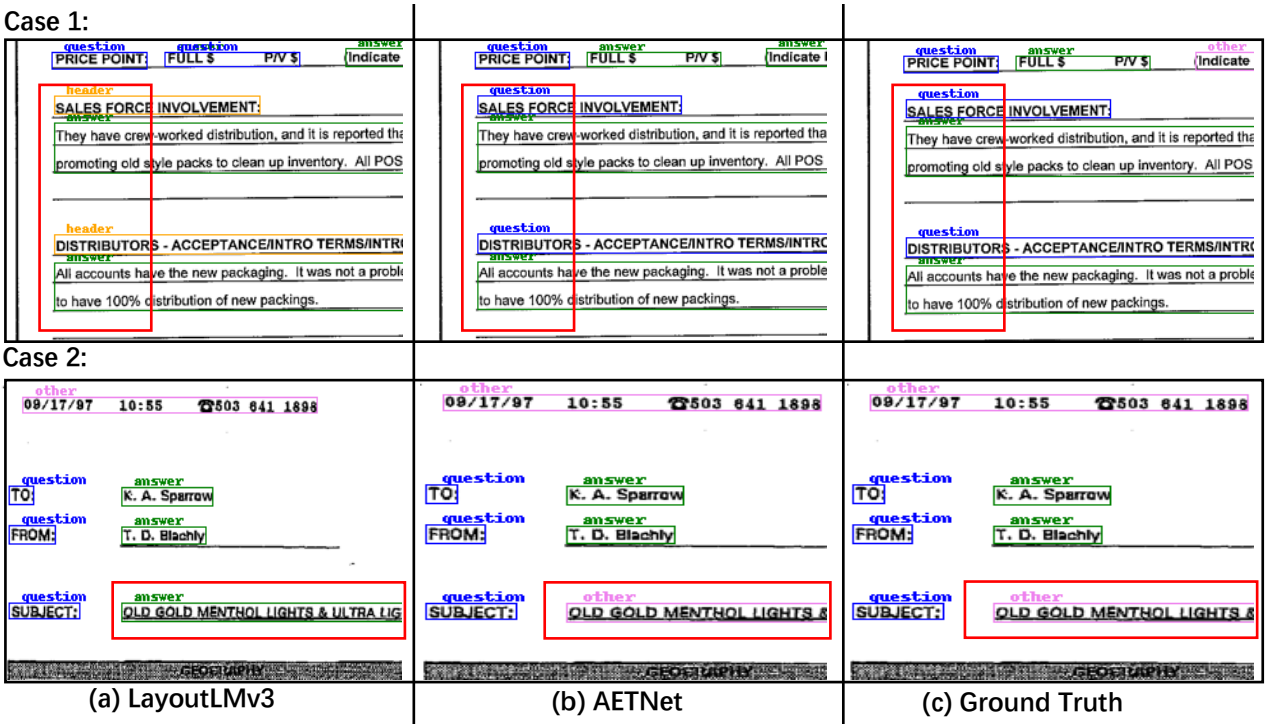

Figure 5: Visualization of two cases on FUNSD, which are predicted by LayoutLMv3 (Column (a)) and AETNet (Column (b). Column (c) is the Ground-truth annotations.

enriched tuning, AETNet achieves the best performance among these three tuning strategies. Specifically, AETNet achieves absolute improvements of $+1.73$ points over fine-tuning and $+7.79$ over prompt-tuning on FUNSD, and $+1.07$ points over fine-tuning and $+9.67$ over prompt-tuning on CORD. It indicates that alignment-enriched tuning has the ability to improve the quality of representations when transferring pre-trained models to downstream tasks.

Ablation Study of each Loss in AETNet To learn the effectiveness of the newly proposed losses (i.e., DITC, IMC, GLITC, and PITA) in improving the document multi-modal representation learning, we perform an ablation study of each each loss in AETNet on FUNSD and CORD datasets. Table 4 report the results. Firstly, it is worth mentioning that AETNet$_{base}$ using supervised objective is slightly worse than LayoutLMv3$_{base}$ using supervised objective. In other words, adding more parameters during the fine-tuning process in this study cannot improve performance. This observation further verifies that downstream tasks benefit from alignment-enriched representations learned by our proposed alignment-enriched objective during the tuning process. Secondly, combining with any newly proposed loss in AETNet$_{base}$ improves the performance on FUNSD and CORD, which can be attributed to the consideration of document-level or patch-level alignment-enriched representations.

Case Study Figure 5 shows the visualization of two cases predicted by LayoutLMv3 and AETNet. In case 1, AETNet predict three entities within the reb box as "question" but LayoutLMv3 predict them as "header". The possible reason is that LayoutLMv3 makes the prediction based on the layout information rather than content text, which requires models to have more accurate interaction and alignment modeling ability to avoid this error. In case 2, LayoutLMv3 learns from two neighbors above it to predict the entity within the red box as "answer", which indicates that the model utilizes too much layout information to make predictions rather than based on understanding it.

Conclusion

This paper proposes a new model architecture with alignment-enriched tuning method named AETNet for transferring multimodal pre-trained models in document images. AETNet consists of an extra visual transformer as the alignment-ware image encoder and an extra text transformer as the alignment-ware text encoder, and a multimodal fusion encoder initialized by pre-trained model parameters. Based on this model architecture, AETNet tunes pre-trained models for downstream tasks with joint task-specific supervised and alignment-aware contrastive objective. AETNet achieves consistent improvements over existing state-of-the-art pre-trained models with fine-tuning and different commonly-used tuning techniques over various downstream tasks, which indicates the effectiveness of our proposed AETNet.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grants (No. 62222203, 61976049 and U20B2063).

References

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. DocFormer: End-to-End Transformer for Document Understanding. In *ICCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387*.

Garncarek, Ł.; Powalski, R.; Stanisławek, T.; Topolski, B.; Halama, P.; Turski, M.; and Graliński, F. 2021. LAMBERT: Layout-Aware Language Modeling for Information Extraction. In *ICDAR*.

Gu, J.; Kuen, J.; Morariu, V.; Zhao, H.; Jain, R.; Barmpalios, N.; Nenkova, A.; and Sun, T. 2021. UniDoc: Unified Pretraining Framework for Document Understanding. In *NeurIPS*.

Gu, Z.; Meng, C.; Wang, K.; Lan, J.; Wang, W.; Gu, M.; and Zhang, L. 2022. XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding. In *CVPR*.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Harley, A. W.; Ufkes, A.; and Derpanis, K. G. 2015. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In *ICDAR*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hong, T.; Kim, D.; Ji, M.; Hwang, W.; Nam, D.; and Park, S. 2022. BROS: A Pre-Trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In *AAAI*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Howard, J.; and Ruder, S. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv preprint arXiv:2204.08387*.

Jaume, G.; Ekenel, H. K.; and Thiran, J.-P. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*.

Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Zhao, T. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; and Park, S. 2022. OCR-Free Document Understanding Transformer. In *European Conference on Computer Vision (ECCV)*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Lee, C.; Cho, K.; and Kang, W. 2019. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*.

Lee, C.-Y.; Li, C.-L.; Dozat, T.; Perot, V.; Su, G.; Hua, N.; Ainslie, J.; Wang, R.; Fujii, Y.; and Pfister, T. 2022. FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction. In *ACL*.

Li, C.; Bi, B.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; and Si, L. 2021a. StructuralLM: Structural Pre-training for Form Understanding. In *ACL*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Li, J.; Xu, Y.; Lv, T.; Cui, L.; Zhang, C.; and Wei, F. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. *arXiv preprint arXiv:2203.02378*.

Li, P.; Gu, J.; Kuen, J.; Morariu, V. I.; Zhao, H.; Jain, R.; Manjunatha, V.; and Liu, H. 2021c. SelfDoc: Self-Supervised Document Representation Learning. In *CVPR*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Qian, Y.; Yu, Y.; Qin, X.; Zhang, C.; Liu, Y.; Yao, K.; Han, J.; Liu, J.; and Ding, E. 2021d. StrucTexT: Structured Text Understanding with Multi-Modal Transformers. In *ACM Multimedia*.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021b. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

Park, S.; Shin, S.; Lee, B.; Lee, J.; Surh, J.; Seo, M.; and Lee, H. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In *Document Intelligence Workshop at Neural Information Processing Systems*.

Powalski, R.; Łukasz Borchmann; Jurkiewicz, D.; Dwojak, T.; Pietruszka, M.; and Pałka, G. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *ICDAR*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, J.; Jin, L.; and Ding, K. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *ACL*.

Wu, T.-L.; Li, C.; Zhang, M.; Chen, T.; Hombaiah, S. A.; and Bendersky, M. 2021. LAMPRET: Layout-Aware Multimodal PreTraining for Document Understanding. *arXiv preprint arXiv:2104.08405*.

Xu, R.; Luo, F.; Zhang, Z.; Tan, C.; Chang, B.; Huang, S.; and Huang, F. 2021a. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; Zhang, M.; and Zhou, L. 2021b. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL*.

Xue, H.; Huang, Y.; Liu, B.; Peng, H.; Fu, J.; Li, H.; and Luo, J. 2021. Probing Inter-modality: Visual Parsing with Self-Attention for Vision-and-Language Pre-training. In *NeurIPS*.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.